Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 20 (2013) 354 – 359

Procedia Computer Science

Complex Adaptive Systems, Publication 3
Cihan H. Dagli, Editor-in-Chief
Conference Organized by Missouri University of Science and Technology
2013- Baltimore, MD.

A Novel Application for Combining CASs and Datasets to Produce Increased Accuracy in Modeling and Predicting Cancer Recurrence

John Norris*, Erin Barns***,Olivia Schultz***, Timothy Masters**, Walker H. Land, Jr.***

** CEO, Health Discovery Corp.(HDC), 531 West Washington Street, Hanson, MA 02341*
*** President, TMAIC, Brackney, PA.*
**** Dept. of BioEngineering, Binghamton Univ., 85 Murry Hill Road, Vestal NY*

Abstract

"Ensemble processing" combines the results (outputs) of several different models, each "looking at" a disease from a different perspective. A number of different methods are available to support ensemble processing: (1) averaging, (2) weighted-averaging, (3) Adaboost, and (4) other processing methods that use gate variables in forming a "tree structure". Gate variables are used here as an integral part of the Expectation operation in a maximum likelihood estimator. This paper presents the application of a "Generalized Regression Neural Network Predictive Model," called the "GRNN oracle," that takes advantage(s) of correlation(s) (synergies) that exist between intelligent predictive input model outputs by combining them (at the variance level) for generating both clinical and microarray lung cancer data to improve cancer recurrence modeling and predictive performance, when compared to any one output taken alone. The hypothesis is: Given a validation data set that contains a **sufficient sample size**, then the GRNN oracle will provide a synergistic combination of output data which is superior in predictive performance accuracy (as measured by an ROC analysis) when compared to all input intelligent models, taken individually. This paper will discuss the results of our work in evaluating the validity of this hypothesis.

Selection and peer-review under responsibility of Missouri University of Science and Technology
Keywords: emsemble processing; mixture of experts; GRNN oracle; statistical learning theory paradigms

1. Introduction

Automated computer aided diagnostic (CAD) systems were first introduced for commercial use in the medical field about a decade ago (R2 [1] and Magic 5[2,3,4] are examples), to be used as a second opinion detection / classification aid by the physician. A few years earlier, ensemble processing (sometimes called "a mixture of experts systems") was also introduced. Ensemble processing had the objective of combining, in an intelligent way, diverse system outputs as inputs to an intelligently combined system, to produce a more intelligent and synergistic output, overall. Ensemble processing combines the results (outputs) of several different models, each looking at a disease from a different perspective. A number of different methods are available to support ensemble processing: (1) averaging, (2) weighted-averaging, (3) Adaboost, and (4) other processing methods that use gate variables to form a "tree structure" rather than as an integral part of a maximum likelihood estimator. A fairly recent review of these candidate methods may be found in the literature (Land, et.al. IJCBDD [5]). The idea behind this processing approach is to produce a synergistic result, one that is a combined and more accurate result than any of the

Selection and peer-review under responsibility of Missouri University of Science and Technology
doi:10.1016/j.procs.2013.09.285

individual model outputs that are then used as combined-system inputs. The objectives of this paper are (1) to combine these two ideas, CAD and ensemble processing, to form an intelligent system, and (2) to use the "generalized regression neural network (GRNN) oracle" (GRNN oracle) model that results to obtain a more accurate estimate for solving medical classification / diagnosis problems. This oracle was developed using the following two hypotheses: First, if two or more intelligent complex adaptive systems (CASs) measure different properties (as characterized by CAS input features) of the same environment, these measurement data contain correlated and synergistic information that can be exploited by an intelligent, adaptive process, which can improve performance. Second, if the method of maximum likelihood combination of these data uses the variance of the estimates (the $(y_i - q_{i,k})^2$ term in eqn.1), an even more accurate synergistic-estimate will be the result. The remainder of this paper contains (1) details on the formulation of the GRNN oracle, (2) some preliminary results of its application to lung cancer data, (3) a discussion of the generated results, (4) the conclusions we reached, and (5) our recommendations for further research and analysis sections to be completed soon or in the months ahead.

2. The GRNN-Based Oracle

An overview of the GRNN-based Oracle is depicted in Figure 1.

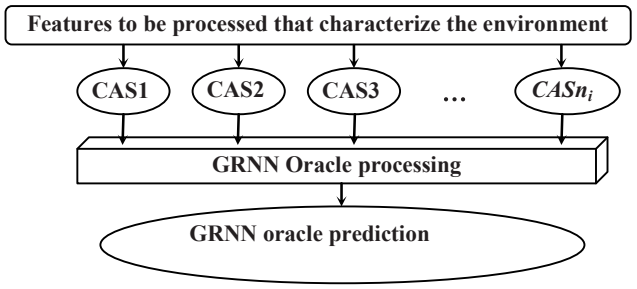

Figure 1. **GRNN Oracle Overview**

Given: two or more prediction models of any type, each of which predicts the same scalar output variable. Neither the nature of these models nor their inputs is important, as the inputs **to** the predictive models are not a part of discussion. (They are not germane to this formulation.) However, one or more *gate variables,* whose values will have an effect on deciding which of the competing models is most "accurate", are important.

The goal here is to design an oracle that uses gate variables to intelligently combine the outputs of competing models, using the following concepts. Once the expected error of each prediction model is estimated, these expected errors mat be used to compute the weights for each model. When an unknown case is processed, the gate variables may be used by the GRNN to decide which CAS outputs are to have the largest weights (i.e, which one operates much like a maximum likelihood estimator). Specifically, certain CAS's are weighted more heavily than the likely inferior CAS's.

For example, for a training set composed of n cases, each case i (i =1,…, n) consists of p gate variables: $x_{i,j}$,where j =1,…, p. These gate variables determine the relative efficacy of the CAS_{ni} prediction models. These m competing prediction models then provide outputs $q_{i,k}$, where k =1,…, m. Now, the **desired output** is y_i. In the following discussion (for simplicity) and for the gate variables and model outputs, just one subscript is used when referring to a trial case that is to be evaluated: x_j where j =1,…, p, are the values of the observed *gate variables*, and q_k where k =1,…, m, are the computed outputs of the m competing prediction models.

Now, a weighted Euclidean distance (as determined by the gate variables) is defined between training case i and the trial case. The GRNN's predicted squared error for model k may be derived to be:

$$\hat{e}_k(\mathbf{x}) = \frac{\sum_{i=1}^{n} (y_i - q_{i,k})^2 \exp(-D_{\mathbf{x}}(\mathbf{x}, \mathbf{x}_i))}{\sum_{i=1}^{n} \exp(-D_{\mathbf{x}}(\mathbf{x}, \mathbf{x}_i))} \quad (1)$$

In order to produce a maximum likelihood estimate, we require that the final prediction be a linear combination of the outputs of the competing models of the following form:

$$\hat{y}(\mathbf{x}) = \sum_{k=1}^{m} w_k q_k \tag{2}$$

We also want the CAS's to have the (desirable) property that their predictions are unbiased. This property is maintained if, and only if, the following condition in equation (3) is imposed:

$$\sum_{k=1}^{m} w_k = 1 \tag{3}$$

Now, the linear combination of unbiased estimators having the maximum likelihood error uses weights proportional to the reciprocal of each estimator's variance. Consequently, when the predicted squared error is used **in place** of the variance in the maximum likelihood variance estimates, the following formula for the weights is derived:

$$w_k = \frac{1/\hat{e}_k}{\sum_{\ell=1}^{m} 1/\hat{e}_\ell} \tag{4}$$

It's important to note that variances (or error expressions) are **expected** to be computed with sufficient accuracy (set experimentally at ±2.5% ,with an estimated ±5% maximum error) for a proper application of this theory. The GRNN oracle is trained (i.e., the p sigma weights are optimized) in the usual cross validation manner. To evaluate the quality of a sigma vector, cases are removed from the data set and the formulas just shown are used first to estimate the competing models' errors, then compute the w_k weights, weight the competing models to get the grand prediction, and finally compute the error of this grand prediction. The sigma vector providing the maximum likelihood estimate is then found. A powerful hybrid training method combining gradient descent and differential evolution has been used to effectively train the sigma values of GRNN oracles (Masters and Land [6]).

3. Preliminary Results With Lung Cancer Data Set

To address the paper's objective, two publicly available data sets, (by Shedden [7] and Raponi [8]), for lung cancer patients, were obtained. The Shedden data set contains 292 patients and Raponi contains 130, and comprise both microarray and clinical data for the same patient set. Using microarray data, from both Raponi and Shedden, a batch-adjustment was performed to compensate for "center-bias" that existed between the two data sets. Then a stratified sample of 20% of the patients was used as a "set-a-side" for a representative set to be used for validation / testing. The 80% clinical data sets, from both Raponi and Shedden, were used for training three CASs, These CASs were: the Probabilistic Neural Networks (PNN), Support Vector Machines (SVM), and Logistic Regression (LR) CASs, and their outputs were used, along with certain specified "gate" variables, as inputs to the GRNN Oracle. Five–fold cross validation was used for validating each of these processes, *which means that 4/5 ^{ths} of the data was used for training and 1/5 for validation. This process is repeated 4 additional times, where a different 4/5^{ths} and 1/5 th of data are held out for training and validation, respectively. Finally, the results of the separate 5 validation set results are averaged for a representative estimate of system performance*. Results of these analyses are depicted in the ROC curves shown in figures 2 and 3 below, which yielded an interesting, but not surprising result: that the GRNN oracle output did not improve the performance, compared to some of the PNN, SVM and LR complex adaptive systems, as it was designed to do. Why? ***We know this is happening because the 5-fold validation sample size is too small, as can be illustrated in the following hypothetical example.*** Using equation (4) of section 2, and assuming that the three CAS system inputs have variances of 2,4 and 8, respectively, a simple evaluation of this equation (4), subject to the constraint imposed by equation (3), shows that these three CAS's inputs combine with weights 57.1%, 28.6%, and 14.3% (which sum to~100%), which is exactly what one would expect in the proper application of the theory described in section 2. That is, the most accurate system inputs receive the highest combination percentage of weight, as specified by the oracle. ***However, section 2 development stipulates that variances are to be computed with a ±2.5% (max. ±5%) experiential error for a proper application of the theory, which brings into play sample size.*** It will be shown in the forthcoming discussion section that these data set sample sizes are ***insufficient*** for an accurate computation of the system variances in a proper application of the GRNN oracle theory. That is, the discussion section will demonstrate that the validation set size is not large enough to result in a ±2.5 % (or at the most a ± 5%) confidence interval for the computed variances.

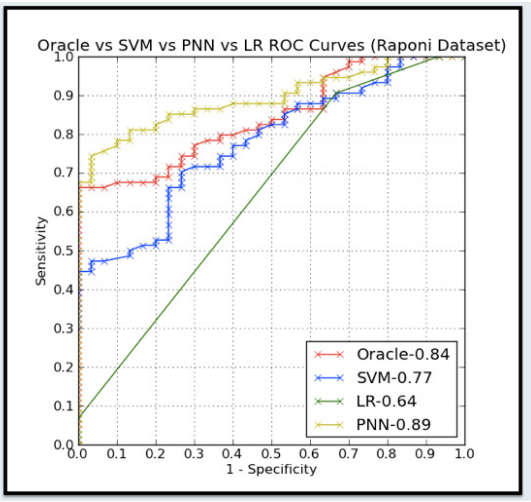

Figure 2. ROC curves for Raponi clinical data set

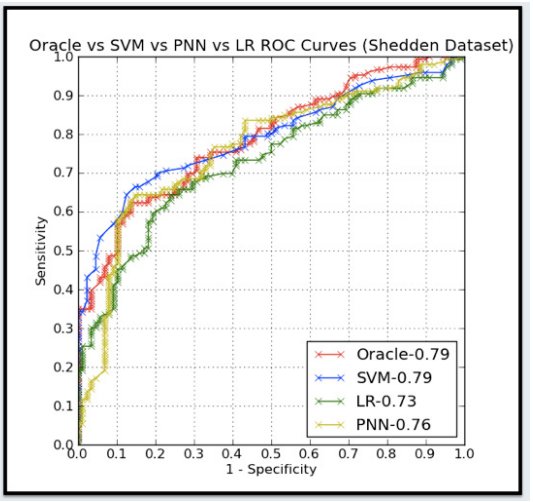

Figure 3. ROC curves for Shedden clinical data set

Simply stated: the sizes (number of patients) for both the Raponi and Shedden clinical data sets are insufficient for a proper application of the GRNN oracle theory. Because these are publically available data sets, the only option is to increase the number of samples in the validation /testing set, which means that the training set size will decrease. That is, a redo of the analysis with two-fold cross validation ***might*** help performance. Such an experiment will decrease the variance confidence interval but is also is expected increase the bias error in the CAS's training phase of the process. Another possible option is to combine the two data sets into one. This analysis was performed. But such a combination only introduced another set of intra- and inter-institutional variability errors (as reflected by a poor AZ) that only exacerbate the result. Another, although remote, possibility is that the physicians at the several cancer research centers, who collected the data, may not have chosen the optimum feature set for best classification accuracy. However, feature selection was out of the hands of the researchers performing this analysis.
The only "correct" option to address this problem is to increase the number of samples (see next section)—not a viable option for this analysis, given the totality of the circumstances. Referring to figure 3, the GRNN Oracle does perform slightly better than the other classifiers for the Shedden clinical data set comparison. This results, as is demonstrated below, because the validation set sample size is larger than that for the Raponi, but still is of insufficient size for the oracle to provide an accurate synergistic estimate. That is, the full range of performance improvement expected from the GRNN Oracle does not occur.

4. Discussion

This discussion section evaluates the GRNN Oracle's performance as a function of data set sizes. ***The following question is addressed: how does the margin of error (i.e., variance of a confidence interval) for a confidence level of 95% change as the test sample size changes?*** This margin of error (as represented by the resulting confidence interval) is a function of the following set of processing parameters: (1) population size (as represented by the respective Raponi and Shedden data set sizes), (2) required sample size (as represented by the validation set size), (3) the confidence interval (which represents how accurately the variance used by the GRNN oracle is computed), and (4) the confidence level (which represents the confidence that the computed variance lies within that confidence interval). These relationships were computed, and then integrated for the 5-fold cross validation process used in this research process,. The results are depicted in figure 4 below. ***The most interesting result from this analysis is that the validation set sizes for the 5-fold cross-validation process are far too low for the GRNN Oracle to be able to provide an accurate result as specified by the theoretical development in section 2***. The GRNN Oracle is designed to work with a 95% confidence interval (or a margin of error of only ±2.5%, or with a maximum of 5%). ***Furthermore, the confidence level of 95% ensures that the oracle variance will be within this confidence interval 95% of the time. Consequently, in order to achieve the ±2.5% margins of error, much larger validation set sizes are required than are available for this analysis, in the given data sets.***

In summary, the margins of error, for the combined Shedden and Raponi data sets, are 11%, 13%, and 20% respectively, when using a 95% confidence level. Specifically for the Raponi validation set samples, we are 95% confident that we are within ±20% of the true value of variance, for the Shedden data set, ±13%. Furthermore, the same interpretations may be applied to the Raponi and Shedden combined validation data set sizes, but with the additional possibility of introducing variability by combining sets that may not be equally representative of the population of interest (even though we attempted to remove the "center bias' in the microarray data). ***These levels of error all exceed the experimentally set ±2.5% confidence interval for variance expected by the oracle.***

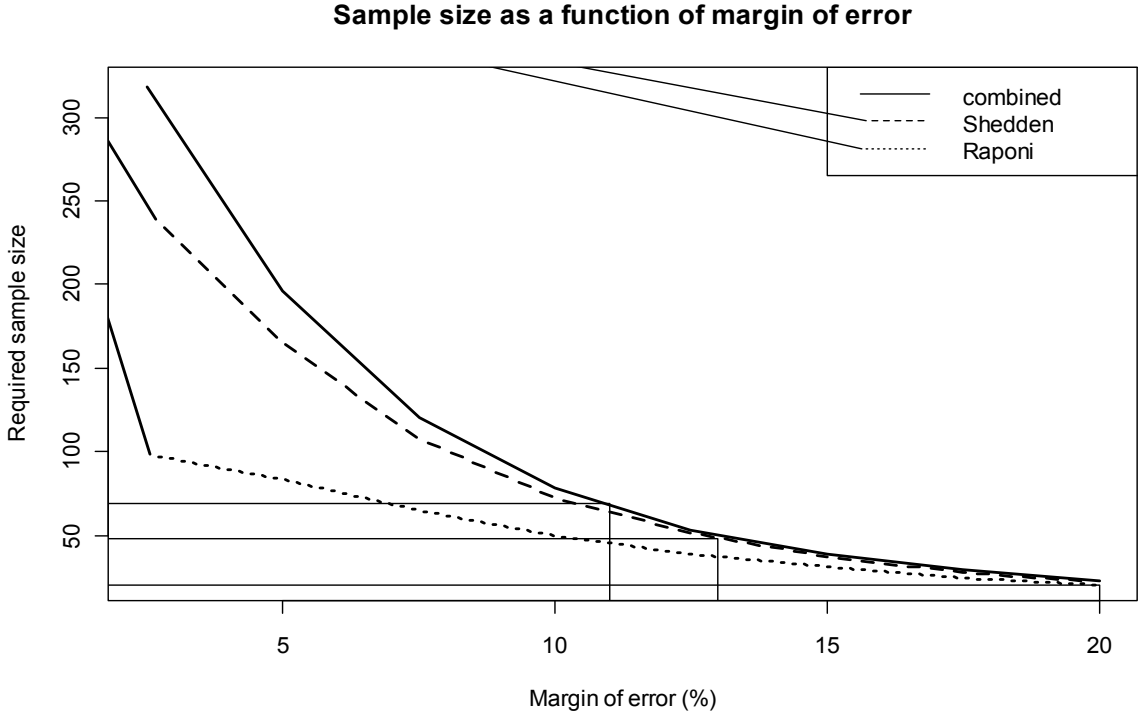

Figure 4. **Confidence Interval (Margin of Error) vs. Required Sample Size**
Graph to ascertain margin of error for a given validation sample set size. The three curves are for the individual Raponi, and Shedden data sets, and the combined data sets.

5. Conclusions

The objective of this paper was to combine intelligent CAS system outputs as inputs for a smart system and use the generalized regression neural network (GRNN) oracle to obtain a more accurate predictive-analytics estimate for solving medical classification / diagnosis problems. The idea was tested using two publically available lung-cancer data sets, referred to in the literature as the Raponi and the Shedden lung-cancer data sets. Three statistical learning theory (SLT) algorithms, PNNs, SVMs, and LR, were used to process these data sets. Then the outputs from these SLT algorithms (CASs) were used as inputs, along with a set of "gate" variables, to a specially designed ensemble processor, with the objective of obtaining a more accurate and synergistic result. However, the result was that the GRNN oracle either did not improve (or only slightly improved) performance of these SLT algorithm processors, a result that was shown to occur because of the inadequate size of the validation set avaliable, for use in a five-fold cross-validation analysis. The only option available for improving the result, we found, would be to increase the amount of lung-cancer data collected, and therefore available for the validation process. This was not a viable option for this research study. The Raponi and the Shedden data sets were also combined, as an option to improve performance. However, this effort, too, was not successful, again because of the insufficient validation data set size,

as well as because of the additional possibilities of errors realized, through introducing unintended center effects.

Throughout the paper we have attempted, repeatedly, to make the point that no matter what tools you have available to you (and use) for data analysis, even SVMs, if the discovery database is not sufficient, in terms of size, but also in terms of content, accuracy, and lack of bias, but especially, most commonly, size, you cannot improve the analysis. The only thing we would add in our future efforts is a further assessment of precisely how a researcher or technician can know when the database is (or databases are) indeed big enough (or not) and when and how he/she can determine that fact before they start the heavy lifting of the analysis. Also, one point not discussed here, due to space limitations, is what impact "targeted pre-processing" of the data, done to assess any bias in the data, might have on how one might optimally proceed before conducting a full analysis. In our next paper, we will assess the importance of all of these additional points.

In sum, we have attempted to craft a thoughtful and thought-provoking article, and to make it clear and very persuasive to readers of diverse backgrounds. But, always, more can be done. The conclusion we reach is of critical importance as the data samples utilized in our analysis are very similar to those used in much of the research going on today in the cancer arena. But data sample sizes are a critical issue that has been dismissed far too casually. Obtaining the right amount of the right type of data (the first time, and every time) is most often the most difficult and expensive part of any new product development effort. Even when multi-factor analysis is an input for SVM (and we use RFE-SVM to eliminate unnecessary data input), two of the most powerful technologies in the world for data analysis for pattern-recognition and machine-learning, SVM can quickly and easily "clog up" and not yield a manageable margin. Getting the database right at the outset, through appropriate forward-thinking (and forward-looking) management, standardization, and, especially, thoughtful pre-processing--all with keen sensitivity to the analysis' specific needs, given the totality of the circumstances-- is critical

6. Recommendations

This research study quantitatively demonstrates the need for using an adequate sample size, for a theory to be accurately evaluated. Consequently, the GRNN oracle we produced should be re-evaluated in coming weeks or months using a data set of sufficient size to properly exercize the theroy we developed in section 2 of this paper.

7. REFERENCES

1.R2 Image checker system (R2 technologies, http://www.r2tech.com)

2.S.Bagnasco, U. Bottigli, P. Cerello, S. Cheran, P. Delogu, M.E. Fantacci,F. Fauci, G. Forni, A. Lauria, E. Lopez Torres, R. Magro, G.L. Masala,P. Oliva R. Palmiero, L. Ramello, G. Raso, A. Retico, M. Sitta, S.Stumbo, S. Tangaro, E. Zanon ": GPCALMA: a GRID based tool formammographic screening, Methods of Information in Medicine 44(2)pp. 244-48, 2005.

3.U. Bottigli, R.Chiarucci, B. Golosio, G.L. Masala, P. Oliva, S.Stumbo,D.Cascio, F. Fauci, M. Glorioso, M. Jacomi, R. Magro and G. Raso ,"Superior Performances of the Neural Network on the Masses Lesions Classification through Morphological Lesion Differences" on IJBSInternational Journal of Biomedical Sciences, Volume 1 Number 1, pp.56-63, 2006.

4.Lauria, R. Massafra, S. Tangaro, R. Bellotti, M. Fantacci, P. Delogu, E. Lopez Torres, P. Cerello, F. Fauci, R. Magro, U. Bottigli,"GPCALMA: an Italian mammographic database of Digitized images for research" proceedings of IWDM, Manchester 18-21Giugno 2006, Lecture Notes in Computer Science 4046, Springer, 2006.

5.Land Jr., W.H., Margolis, D., Kallergi, M. and Heine, J.J. "A kernel approach for ensemble decision combinations with two-view mammography applications", Int. J. Functional Informatics and Personalised Medicine, Vol. 3, No. 2,pp.157–182, 2010.

6.T. Masters and W.H. Land, Jr., "A new training method for the General Regression Neural Network," IEEE International SMC Conference Proceedings, pp. 1990-5, 1997.

7.Kerby Shedden, Jeremy M G Taylor, Steven A Enkemann, Ming-Sound Tsao, Timothy J Yeatman, William L Gerald, et. al., Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study, Nature Medicine, Vol. 14, No. 8 (2008) 822-827.

8. *Raponi M, Zhang Y, Yu J, Chen G, Lee G, Taylor JMG, Macdonald J, Thomas D, Moskaluk C, Wang Y, Beer DG: Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer research 2006, 66(15):7466-72[http://cancerres.aacrjournals.org/ cgi/content/abstract/66/15/7466].*